Bayesian Hierarchical modeling and Inference for Mechanistic Systems in Industrial Hygiene

Soumyakanti Pan
UCLA Department of Biostatistics,
University of California Los Angeles,
Los Angeles, California 90095-1772, USA
span18@ucla.edu

Darpan Das
Department of Environment and Geography,
University of York,
Heslington, York, Y010 5NG, UK
darpan.das@york.ac.uk

Gurumurthy Ramachandran
Department of Environmental Health Sciences and Engineering,
Johns Hopkins Bloomberg School of Public Health and Whitmore School of Engineering,
Baltimore, Maryland, 21205, USA
gramach5@jhu.edu

Sudipto Banerjee
UCLA Department of Biostatistics,
University of California Los Angeles,
Los Angeles, California 90095-1772, USA
sudipto@ucla.edu

July 4, 2023

Abstract

A series of experiments in stationary and moving passenger rail cars were conducted to measure removal rates of particles in the size ranges of SARS-CoV-2 viral aerosols, and the air changes per hour provided by existing and modified air handling systems. Such methods for exposure assessments are customarily based on *mechanistic* models derived from physical laws of particle movement that are deterministic and do not account for measurement errors inherent in data collection. The resulting analysis compromises on reliably learning about mechanistic factors such as ventilation rates, aerosol generation rates and filtration efficiencies from field measurements. This manuscript develops a Bayesian state space modeling framework that synthesizes information from the mechanistic system as well as the field data. We derive a stochastic model from finite difference approximations of differential equations explaining particle concentrations. Our inferential framework trains the mechanistic system using the field measurements from the chamber experiments and delivers reliable estimates of the underlying physical process with fully model-based uncertainty quantification. Our application falls within the realm of Bayesian "melding" of mechanistic and statistical models and is of significant relevance to environmental hygienists and public health researchers working on assessing performance of aerosol removal rates for rail car fleets.

Keywords: Bayesian inference; Dynamical systems; Industrial Hygiene; Mechanistic systems; Melding; Differential equations; State-space models

1 Introduction

With the outbreak of the Covid-19 pandemic, public transit demand in the United States took a hit (NYC MTA, 2020) as initial reports suggested it to be among the major vectors for transmission of the Sars-Cov-2 virus (Harris, 2020). As it became clearer that the virus causing COVID-19 was transmitted via respiratory secretions which are aerosolized into tiny droplets (Chia *and others*, 2020), transit agencies took measures to reduce the exposure to SARS-CoV-2 and the probability of infection for passengers and employees. Following studies revealing inadequate social distancing rules in such settings (Bazant and Bush, 2021), transit agencies have considered engineering interventions with the aim of reducing the risk of infection. While ventilation and filtration have always been integral to the air handling systems of train fleets, the COVID-19 public health crisis has brought increased attention on the effectiveness of engineering interventions.

In partnership with a large-scale, interstate, mass-transit rail company in the US, researchers have carried out a series of experiments inside a fleet of passenger rail cars

sampled with a design accounting for various controls involving ventilation and filtration systems. The experiments focus on measuring concentration of aerosols at different locations inside the rail car compartment with an aerosol generator in the center. The aim is to ascertain important quantities related to ventilation and filtration system. As we do not observe the actual aerosol concentrations directly, but record partial noisy measurements, it is crucial from the environmental hygienist's perspective to understand the underlying physical process described by a system of deterministic differential equations.

Consolidating scientific inference by borrowing information from deterministic mechanistic systems and from field measurements designed to emulate the system continues to attract significant attention in diverse health science applications. Statistical approaches include Bayesian melding (e.g., Raftery *and others*, 1995; Poole and Raftery, 2000; Fuentes and Raftery, 2005; Raftery and Bao, 2010), which synthesizes such information through a generic Bayesian hierarchical framework,

$$[\text{data} \,|\, \text{process}, \text{parameters}] \times [\text{process} \,|\, \text{parameters}] \times [\text{parameters}] \,. \tag{1}$$

by modeling the field measurements (data), the mechanistic system (process) and all model parameters (mechanistic and statistical) jointly using probability distributions. Bayesian inference typically computes, or draws samples from, the posterior distribution of the process and parameters and carries out subsequent predictive inference by extending such inference to hitherto unmeasured observations. In its simplest form, Bayesian melding proceeds by regressing the data on the physical model. See, for example, Zhang *and others* (2009) and Raftery and Bao (2010) for two different applications. Monteiro *and others* (2014) demonstrate, however, that straightforward Bayesian nonlinear regression can be highly ineffective in predicting exposure concentrations in designed chamber experiments such as those encountered here.

Using stochastic process emulators to model the output of the mechanistic system is widely used in calibrating computer models and similar approaches have been used in Bayesian melding (see, e.g., Monteiro *and others*, 2014; Fuentes and Raftery, 2005). In fact, such methods are often the only option when the mechanistic system is highly complex (e.g., climate models) and requires very specialized computing environments for implementation. In industrial hygiene, on the other hand, relatively simple differential equations comprise the mechanistic system which suggests building Bayesian dynamical systems for their analysis (Abdalla *and others*, 2020; Wikle and Hooten, 2010; Wikle *and others*, 2019). This allows the mechanistic parameters to directly learn from the data obviating the need to carefully design runs, often multiple times, of the mechanistic system over a range of inputs. We work within such a paradigm here.

The novelty of our application lies in the manner in which we address several data analytic challenges. First, the mechanistic models we consider incorporate multiple rise

and decay of concentrations that are governed by the mechanistic parameters and experimental conditions. Assimilating this information requires a careful balance of statistical learning from the data as well as from the underlying deterministic mechanism. Second, we need to construct our inferential framework to handle streaming in as different cycles within the experiment. Industrial hygiene experiments typically involve a substantial amount of unreliable "background data" between cycles. We address this issue by allowing our framework to learn about the process in these background zones by assimilating mechanistic considerations with data driven inference. A specific contribution of this framework is aimed at public health researchers as we show the inferential benefits of performing an analysis by delving into the mechanistic equations over a black-box emulator-based inference based on multiple runs of the system.

The remainder of this manuscript evolves as follows. Section 2 offers an account of different mechanistic models in industrial hygiene and offers scientific justification for our framework. Section 3 describes the design and conduct of the field experiment. Section 4 develops the Bayesian hierarchical modeling framework while Sections 5 and 6 present analysis of simulated data and that of the field experiment, respectively. Section 7 concludes the article with a discussion.

2 Mechanistic Models

The "one box model" (Reinke and Keil, 2009) is widely used in environmental engineering to assess occupational exposure when subject exposure occurs far from the source. The working assumptions of the model includes the "well-mixed room" assumption indicating a spatial uniformity of particle concentration inside the chamber at an instant. The assumption of the room being well mixed is due to either natural or induced air currents, which results in nearly equal concentration levels throughout the room. However, the standard well-mixed room model in presence of local controls and modifications are needed (Hewett and Ganser, 2017).

The standard model assumes that a source is generating a pollutant at a constant rate G in a room of volume V and ventilation volumetric flow rate Q. The following differential equation describes the dynamics of particle concentration $C(.)$ inside the room, which is a function of time t. We will refer to this system as "Model 101" (acronym 1Box.CE.Gv in Hewett and Ganser, 2017).

$$\text{Model 101:} \quad V\frac{dC}{dt} = G - CQ \tag{2}$$

Using the initial particle concentration of the room, C_0, we can find the following closed

form solution to (2) describing the time dependent particle concentrations,

$$C(t; C_0, \phi) = C_0 \exp\left(-\frac{Q}{V}t\right) + \frac{G}{Q}\left[1 - \exp\left(-\frac{Q}{V}t\right)\right] , \tag{3}$$

where $\phi = (G, Q)$ denotes unknown parameters of interest and V is known from the specifications of the chamber experiment. The following systems indicates that the particle concentration under constant emission G and constant ventilation Q will reach a steady state concentration of approximately $\lim_{t\to\infty} C(t) = G/Q$. Usually, the generation is stopped after some time and the concentration eventually decays resulting in an experiment cycle. If the total time taken by an experiment cycle to end is T with the generation stopped at time T_0, then the time dependent concentration during the exposure rise and decay of a cyclic process is given by the functions C_r and C_d as in (4) and (5).

$$\text{Rise: } C_\mathrm{r}(t; C_0, \phi) = C_0 \exp\left(-\frac{Q}{V}t\right) + \frac{G}{Q}\left[1 - \exp\left(-\frac{Q}{V}t\right)\right], \qquad t \le T_0 \tag{4}$$

$$\text{Decay: } C_\mathrm{d}(t; C_0, \phi) = C_\mathrm{r}(T_0; C_0, \phi) \exp\left(-\frac{Q}{V}(t - T_0)\right), \qquad t > T_0 \tag{5}$$

Due to the inadequacy of Model 101 for exposure assessment in the presence of local engineering controls, Hewett and Ganser (2017) propose enriching the model with suitable parameters for local controls and develop a nested sequence of mechanistic models. The last, hence the richest, model in the sequence is described as "one box, constant emissions, Local Exhaust Ventilation (LEV) with return, general ventilation with recirculation" (acronym 1Box.CE.LevR.GvR). This model is applicable to a local exhaust setting in which the filtered air is returned to the workplace, but with an increase in the effective ventilation by the amount of recirculated air, accompanied by the efficiencies for contaminant collection, filtration and general ventilation re-circulation. We refer to this model as "Model 111", which is described by the mass balance equation,

$$\text{Model 111:} \quad V\frac{dC}{dt} = (1 - \epsilon_L\epsilon_{L.F})G - C(Q + \epsilon_{L.F}Q_L + \epsilon_{R.F}Q_R) . \tag{6}$$

The closed form solution of (6) is a reparametrized version of the functions in (4) and (5), $C_\mathrm{r}(t; \phi')$ and $C_\mathrm{d}(t; \phi')$ with $\phi' = (G', Q')$ where, $G' = (1-\epsilon_L\epsilon_{L.F})G$ and $Q' = Q+\epsilon_{L.F}Q_L+\epsilon_{R.F}Q_R$. The parameters of interest are $\phi_1 = \{G, Q, Q_R, Q_L, \epsilon_L, \epsilon_{L.F}, \epsilon_{R.F}\}$. Table 1 briefly explains the parameters involved in this dynamic system.

3 Experiment

Experimental investigations were carried out on flow rate in three rail cars of the same fleet, representative of the rail company's most regularly used commuter passenger cars.

Variable	Definition	Unit
G	Generation rate	mg/min
V	Volume	m^3
Q	Ventilation rate	m^3/min
Q_L	local exhaust ventilation rate	m^3/min
Q_R	room recirculation system ventilation rate	m^3/min
ϵ_L	fraction of the source emissions immediately captured by the local exhaust	unitless (0,1)
$\epsilon_{L.F}$	local exhaust return filtration efficiency	unitless (0,1)
$\epsilon_{R.F}$	general ventilation recirculation filtration efficiency	unitless (0,1)

Table 1: Hewett Model 111 parameters

Each rail car was 150.5 m^3 (5,314 ft^3) with a designed outdoor air intake flow rate of 34 m^3/min and a designed total supply air flow rate of 102 m^3/min. The air in the car is designed to be filtered 40.7 times per hour and replaced or changed with outdoor air 13.6 times per hour by the HVAC system. Outdoor air is brought into the rail cars' return air duct (return plenum) through dampers that regulate the airflow. Here, the outdoor air mixes with the recirculated air, passes through a MERV-8/13 filter,then moves through the heating and cooling elements before entering the supply air duct (supply plenum) to be distributed back into the car volume. An exhaust blower removes a portion of the cabin air to the outside depending on the position of a ventilation damper.

The rail cars can operate at speeds up to 201 km/h (125 mph). Each cabin had 36 seats on each side of a central aisle, spread over 18 rows, overhead compartments above each row, and two bathrooms on one end as shown in Figure 2. Aerosols in the 0.3–5.0 mm size range were generated using a Collision nebulizer (MRE 3-jet with attached pressure gauge) with a 70:30 mixture of propylene glycol and vegetable glycerin. The nebulizer was placed in the center of the rail car between rows 10 and 11 (Figure 2), on a stand 1.0 m above the floor with the outlet 0.2 m above that. This height is equivalent to the distance from the floor to the middle part of the seat's headrest, making it a good approximation for the height of a person's breathing zone and the origin of particle dispersion.

Real-time aerosol concentrations were measured at four locations in the passenger cars using photo detector particle counters (AeroTrak Handheld Particle Counter- Model 9306; TSI; Shoreview, MN). The AeroTrak counts particles using a laser beam and a photodetector to detect light scattering and provides particle counts in six size ranges: 0.3–0.5 mm, 0.5–1.0 mm, 1.0–3.0 mm, 3.0–5.0 mm, 5.0–10.0 mm, and $>$ 10.0 mm. Each AeroTrak was calibrated daily, before beginning the experiments. Aerosol concentration measurements were logged at 1-min intervals for each experiment and downloaded to a computer as .csv files. Each experimental run consisted of 3 experiment cycles with

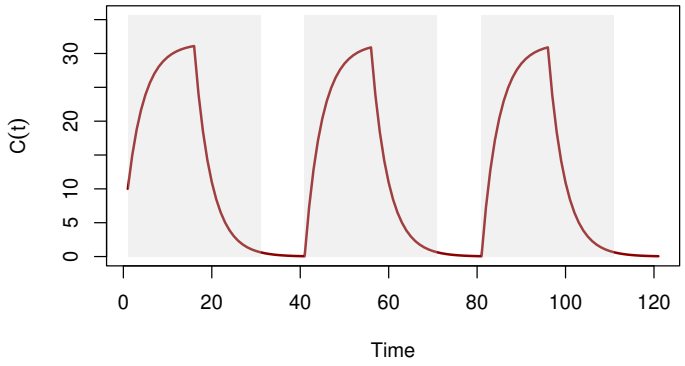


Figure 1: Plot of concentration curve $C(t)$ versus t for a cyclic experiment with 3 cycles. Total length of the experiment is 80 minutes, with emission occurring during the first 15 minutes of each cycle. The plot uses the following test values of the model parameters: $G = 1000$ mg/min, $V = 100$ m^3, $Q = 20$ m^3/min, $Q_L = 5$ m^3/min, $Q_R = 5$ m^3/min, $\epsilon_L = 0.6$, $\epsilon_{LF} = 0.3$, $\epsilon_{RF} = 0.9$, $C_0 = 10$ mg. Occasionally, we analyze equal time intervals of exposure rise and decay (here, it is 15 minutes each) for a balanced design as shown in the plot by the area shaded in gray with the remaining data treated as "background".

each cycle carried out over a period of approximately 30 min with some background at the end, with the Collison nebulizer generating the aerosol for the first 15 min (aerosol concentration increase) and no aerosol generation for the second 15 min (aerosol concentration decrease). The intent was not to mimic human breathing or speaking but rather to observe the fate of aerosol particles of relevant sizes over time in the cabin. Complete details of the sampling instrumentation and experimental design are given in Das *and others* (2023). The Hewett model 111 as stated in (6) reasonably conforms with the experimental setup described above.

4 Bayesian modeling

The statistical model must account for the considerable amount of measurement errors and suitably quantify uncertainties in the field experiment. Wikle and Hooten (2010) offer a broad framework for statistical modeling exploiting knowledge of the underlying physical system available in the form of a dynamical system. We assume that a first-order Markov assumption is appropriate in this context and, hence, we introduce a process evolution model describing the latent true particle concentrations inside the chamber.

The basic framework follows (7). Due to a high degree of skewness in particle concentrations, it is reasonable to model the logarithmic concentration with Gaussian noise. Let Y_t denote the measured concentration at time t and let C_t be the latent process representing the true concentration at time t. The observation equation allows the la-

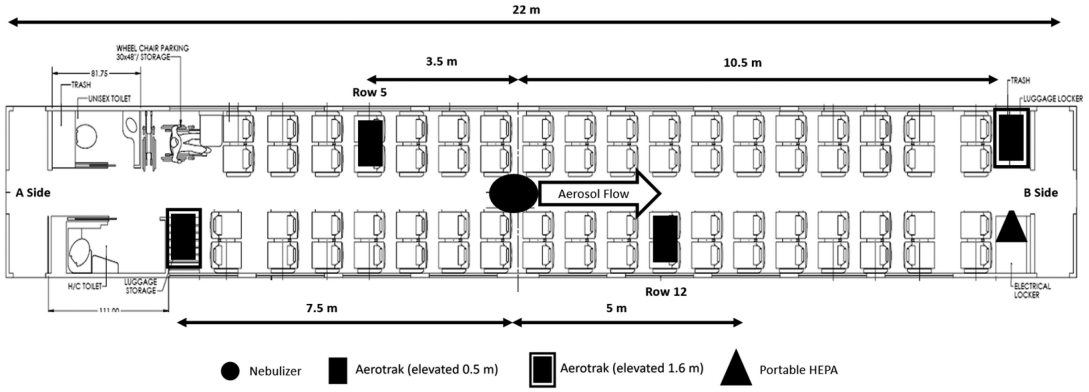


Figure 2: A schematic diagram of the experimental setup in a typical passenger car, drawn to scale with lengths in metres.

tent concentration to drive the inference while accommodating measurement errors. The transition equation models concentrations over time. These are formulated as

$$\begin{aligned} \text{Observation Equation:} \quad & \log Y_t = \log C_t + \upsilon_t, \ \upsilon_t \sim P_\upsilon \\ \text{Transition Equation:} \quad & C_t = f(C_{t-1}) + \omega_t, \ \omega_t \sim P_\omega \ , \end{aligned} \tag{7}$$

where υ_t and ω_t are random processes modeling measurement errors and uncertainty in the concentration process through probability distributions P_υ and P_ω, respectively, and $f(\cdot)$ is a specified function to introduce non-linearity in the transitions if needed.

Replacing the instantaneous rate of change of concentration in (2) by the average change in concentration in a time interval $(t, t + \Delta_t]$, yields an approximate relation between the concentration at the end and at the beginning of the interval. If $C(t + \Delta_t)$ is the underlying particle concentration at the next time point of measurements, with Δ_t being specified according to the time gap between successive measurements during the experiment and units of relevant parameters, we model the rise and decay as $C_{t+1} \approx \left(1 - \frac{\Delta_t}{V}Q\right) C_t + \frac{\Delta_t}{V}G$ and $C_{t+1} \approx \left(1 - \frac{\Delta_t}{V}Q\right) C_t$, respectively. Therefore,

$$\text{Observation:} \quad \log Y_t = \log C_t + X_t^\top \beta + \upsilon_t, \ \upsilon_t \sim P_\upsilon \tag{8}$$

$$\text{Transition:} \quad C_t = \left(1 - \frac{\Delta_t}{V}Q\right) C_{t-1} + \frac{\Delta_t}{V}G_t + \omega_t, \ \omega_t \sim P_\omega \tag{9}$$

where, X_t is a vector of explanatory variables at time t, $G_t = G1_{\mathcal{G}}(t)$, $1_{\mathcal{G}}(t)$ is the indicator function for $t \in \mathcal{G}$ and $\mathcal{G}$ is the collection of time points when the generation of particles was in place. The random process υ_t accounts for observation error and ω_t accounts for errors originating from the finite difference approximation of the differential equation and for possible biases in the deterministic model.

4.1 Hierarchical Bayesian State-Space Model

We present a Bayesian state space model derived from (8) and (9) that address two challenges. First, the nature of the experiment generates consecutive cycles of data as described in Figure 1. Second, each cycle is composed of both a rise and decay in concentrations as described in (4), where the initial concentration of a cycle is derived from the estimated concentration in the second cycle. If Z_t is the (possibly transformed) observed data and $g(\cdot)$ is a suitable transformation for the latent particle concentration at time t, C_t, then we construct the Bayesian dynamic model

$$\begin{aligned} Z_t &= g(C_t) + X_t^\top \beta + \upsilon_t,\ \upsilon_t \overset{\text{iid}}{\sim} P_{\tau_1}, \\ C_t &= A_t(\phi, \Delta_t)\, C_{t-1} + B_t(\phi, \Delta_t) + \omega_t,\ \omega_t \overset{\text{iid}}{\sim} P_{\tau_2}, \\ \{\phi, \beta, \tau\} \mid \psi &\sim p(\phi)\, p(\beta, \tau \mid \psi), \\ \{\psi\} &\sim \pi(\psi) \end{aligned} \tag{10}$$

where, $\tau = \{\tau_1, \tau_2\}$ are the parameters associated with the error distributions. The coefficients A_t and B_t in the process evolution are functions of ϕ, the unknown parameters of the mechanistic model and the finite difference increments Δ_t, which are known. In Model 101, $\phi = \{G, Q\}$ whereas, in Model 111, $\phi = \{G, Q, Q_R, Q_L, \epsilon_L, \epsilon_{L.F}, \epsilon_{R.F}\}$. Usually prior information on the such parameters is scarce and, hence, uniform priors are considered. As we are modeling particle concentrations, it is reasonable to consider $g(\cdot)$ as the logarithm function and P_{τ_1} as the Gaussian distribution.

Since the process evolution models particle concentration, we restrict P_{τ_2} to a distribution with non-negative support. A log-normal distribution for P_{τ_2} possibly dependent on time t is a viable choice. Abdalla *and others* (2020) have used the Gamma distribution in mechanistic settings. Letting $Z_t = \log Y_t$, where Y_t are the observed concentrations, we consider the following model incorporating mechanistic Model 111 in (6),

$$\begin{aligned} Z_t \mid C_t, \beta, \sigma_\upsilon^2 &\sim \mathcal{N}(\log C_t + X_t^\top \beta, \sigma_\upsilon^2) \\ C_t \mid \phi, m_\omega, \sigma_\omega^2 &\sim \text{ShiftedLN}(A_t(\phi, \Delta_t)\, C_{t-1} + B_t(\phi, \Delta_t); m_\omega, \sigma_\omega^2) \\ \{\phi, \beta, \sigma_\upsilon^2, m_\omega, \sigma_\omega^2\} &\sim p(\phi)\, p(\beta \mid \sigma_\upsilon^2)\, p(\sigma_\upsilon^2)\, p(m_\omega)\, p(\sigma_\omega^2)\ , \end{aligned} \tag{11}$$

where $A_t(\phi, \Delta_t) = 1 - (Q + \epsilon_{L.F} Q_L + \epsilon_{R.F} Q_R)\Delta_t / V$ and $B_t(\phi) = G_t \Delta_t / V$ with $G_t = (1 - \epsilon_L \epsilon_{L.F}) G 1_{\mathcal{G}}(t)$ as described in (9). The random variable $X + \theta$ is said to be distributed as shifted log-normal $\text{ShiftedLN}(\theta; \mu, \sigma^2)$ if $\log X$ is distributed normally with mean μ and variance σ^2 for some $\theta \in \mathbb{R}$. Setting $Q_L = Q_R = \epsilon_L = 0$ in (11) obtains a hierarchical model for Model 101 (2).

4.2 Model for observed and latent states

A salient feature of our analysis concerns the experiment being composed of K cyclic experiments over the time period $\mathcal{T} = [0, T]$ with measurements taken over an ordered set of time points $0 < t_1 < t_2 < \cdots < t_N$, where N is the total total number of observed time points. Recognizing that the background data (see Section 2 and, more specifically, Figure 1) collected between two cyclic experiments are often deemed unreliable, we estimate, with uncertainty quantification, the concentration state at the end of a cycle and use it as the assumed value at the start of of next cycle. We use the Bayesian hierarchical model in (10) to jointly model the observations and latent states over all the cycles.

Let $\mathcal{K} = \{t_1, t_2, \ldots, t_N\}$ be the set of time points at which the concentrations are measured over the duration of the experiment. We partition $\mathcal{K} = \sqcup_{i=1}^{K} \mathcal{K}_i$ into K distinct cycles, where $\mathcal{K}_i$ denotes all the time points generating measurements in cycle $i \in \{1, 2, \ldots, K\}$ of the experiment and $\sqcup$ denotes disjoint unions. Let $t_i = \max \mathcal{K}_i$ be the last time point measuring concentrations for cycle i. Let $Y_{\mathcal{K}} = \{Y_{t_j} : t_j \in \mathcal{K}\}$ and $C_{\mathcal{K}} = \{C_{t_j} : t_j \in \mathcal{K}\}$ denote the sets of measurements and latent states of concentrations, respectively. The parameter space is given by $\Theta = \Theta_1 \sqcup \Theta_2$, where Θ_1 and Θ_2 are parameters present in the observation and latent equations, respectively.

Building a hierarchical stochastic model for the observations and latent states conforming to (11) will need to account for the latent state at the end of a cycle as the value of the concentration state at the start of the next cycle is learned from the former. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, where s_i denotes the starting time point of cycle i. We note that s_i signifies the start of cycle i and, therefore, is possibly distinct from the first time point in $\mathcal{K}_i$, which is the time point for the first measurement in cycle i. Therefore, $s_i \leq \min \mathcal{K}_i$. Assuming that the cycles are conditionally independent, given Θ, the joint distribution of $Y_{\mathcal{K}}$ and $C_{\mathcal{K}}$ is

$$p(Y_{\mathcal{K}}, C_{\mathcal{K} \cup \mathcal{S}} \mid \Theta) = \prod_{i=1}^{K} p(C_{s_i} \mid C_{u_{i-1}}, \Theta_2) \prod_{t_j \in \mathcal{K}_i} p(Y_{t_j} \mid C_{t_j}, \Theta_1) p(C_{t_j} \mid C_{t_{j-1}}, \Theta_2) \,, \tag{12}$$

where $u_i = \max \mathcal{K}_i$ denotes the end point of cycle i and $p(C_{s_1} \mid C_{u_0}, \Theta_2) = p(C_{s_1})$, which quantifies belief about the concentration state at the beginning of the first cycle, hence the starting condition of the experiment itself.

The distributions $p(C_{t_j} \,|\, C_{t_{j-1}}, \Theta_2)$ and $p(Y_{t_j} \,|\, C_{t_j}, \Theta_1)$ are specified as shifted log-normal and log-normal, respectively, as in (11). The parameters in (5) appear in (12) as $\Theta_1 = \{\beta, \sigma_v^2\}$ and $\Theta_2 = \{G, Q, Q_R, Q_L, \epsilon_L, \epsilon_{L.F}, \epsilon_{R.F}, m_\omega, \sigma_\omega^2\}$. We assign a log-normal distribution for $p(C_{s_i} \,|\, C_{u_{i-1}}, \Theta_2)$ such that $\log C_{s_i} \sim \mathcal{N}(\log \mu_{s_i}, \sigma_\omega^2)$, where $\log(\mu_{s_i}) = \log C_{u_{i-1}} - (Q/V)(s_i - u_{i-1})$ is derived from the mechanistic considerations embodied in (5). Therefore, the latent concentration state at the beginning of a cycle learns from mechanistic considerations while also accounting for dispersion using the log-normal dis-

tribution. These specifications ensure a dynamic framework even as we marginalize over $C_{\mathcal{S}}$ leaving the distribution of the observed data dependent only on Θ. Hence, for a fixed initial concentration $C_{s_1} = C_0$, (12) yields the joint distribution

$$\begin{aligned} p(Y_{\mathcal{K}}, C_{\mathcal{K}} \mid \Theta) &= \int \prod_{i=1}^{K} p(C_{s_i} \mid C_{u_{i-1}}, \Theta_2) \prod_{t_j \in \mathcal{K}_i} p(Y_{t_j} \mid C_{t_j}, \Theta_1) p(C_{t_j} \mid C_{t_{j-1}}, \Theta_2) dC_{\mathcal{S}} \\ &= \prod_{t_j \in \mathcal{K}} p(Y_{t_j} \mid C_{t_j}, \Theta_1) p(C_{t_j} \mid C_{t_{j-1}}, \Theta_2) \ . \end{aligned} \tag{13}$$

This reveals that the Markovian dependence within a cycle $\mathcal{K}_i$ in (12) is retained for any time point in $\mathcal{K}$.

4.3 Prior and Posterior

We extend (12) to a joint distribution for $\{\Theta, C_{\mathcal{K}\cup\mathcal{S}}, Y_{\mathcal{K}}\}$ by specifying a prior distribution $p(\Theta)$. The posterior distribution is proportional to the joint distribution

$$p(\Theta, C_{\mathcal{K}\cup\mathcal{S}} \,|\, Y_{\mathcal{K}}) \propto p(\Theta) \prod_{i=1}^{K} p(C_{s_i} \mid C_{s_{i-1}}, \Theta_2) \prod_{t \in \mathcal{K}_i} p(Y_t \mid C_t, \Theta_1) p(C_t \mid C_{t-1}, \Theta_2) \ , \tag{14}$$

where the prior distribution corresponding to (11) is given by

$$\begin{aligned} p(\Theta) = \mathcal{N}\left(\beta \,|\, \mu_\beta, \alpha\sigma_v^2\right) &\times \mathcal{IG}\left(\sigma_v^2 \,|\, a_v, b_v\right) \times \mathcal{IG}\left(\sigma_\omega^2 \,|\, a_\omega, b_\omega\right) \times \mathcal{N}(m_\omega \,|\, \mu_m, \kappa_m) \\ &\times \mathcal{U}\left(G \,|\, a_G, b_G\right) \times \mathcal{U}\left(Q \,|\, a_Q, b_Q\right) \times \mathcal{U}\left(Q_L \,|\, a_{Q_L}, b_{Q_L}\right) \times \mathcal{U}\left(Q_R \,|\, a_{Q_R}, b_{Q_R}\right) \\ &\times \mathcal{U}\left(\epsilon_L \,|\, a_{\epsilon_L}, b_{\epsilon_L}\right) \times \mathcal{U}\left(\epsilon_{L.F} \,|\, a_{\epsilon_{L.F}}, b_{\epsilon_{L.F}}\right) \times \mathcal{U}\left(\epsilon_{R.F} \,|\, a_{\epsilon_{R.F}}, b_{\epsilon_{R.F}}\right) \ , \end{aligned} \tag{15}$$

where we denote $\mathcal{N}(X \,|\, a, b)$, $\mathcal{IG}(X \,|\, a, b)$ and $\mathcal{U}(X \,|\, a, b)$ as Normal, inverse-Gamma and Uniform densities in X with parameters a and b, respectively (Gelman *and others*, 2013).

4.4 Smoothing

A key inferential objective in dynamical systems is the smoothing of the latent process generating the data. In our current context, this amounts to model-based inference for the values of the latent concentrations at unobserved time points. Let $\mathcal{Z}$ be a finite collection of arbitrary time points where concentrations have not been measured. These points can be situated within the time duration of a cycle, a background time point for a cycle, or as a future time point of a cycle.

We use the posterior distribution $p(\Theta, C_{\mathcal{K}} \,|\, Y_{\mathcal{K}})$ to evaluate the predictive distribution

$$p(C_{\mathcal{Z}} \mid Y_{\mathcal{K}}) = \int p(C_{\mathcal{Z}} \mid Y_{\mathcal{K}}, C_{\mathcal{K}}, \Theta)\, p(\Theta, C_{\mathcal{K}} \mid Y_{\mathcal{K}})\, d\Theta\, dC_{\mathcal{K}} \ . \tag{16}$$

Sampling from (16) is achieved as follows. For each value of $\{\Theta, C_{\mathcal{K}}\}$ sampled from $p(\Theta, C_{\mathcal{K}} \,|\, Y_{\mathcal{K}})$, we draw one sample of $C_{\mathcal{Z}}$ from the conditional predictive distribution $p(C_{\mathcal{Z}} \mid Y_{\mathcal{K}}, C_{\mathcal{K}}, \Theta)$. Furthermore, we sample from the posterior predictive distribution of the measurements

$$\begin{aligned} p(Y_{\mathcal{Z}} \mid Y_{\mathcal{K}}) &= \int p(Y_{\mathcal{Z}} \mid Y_{\mathcal{K}}, C_{\mathcal{K}}, \Theta)\, p(C_{\mathcal{K}}, \Theta \mid Y_{\mathcal{K}})\, d\Theta\, dC_{\mathcal{K}} \\ &= \int p(Y_{\mathcal{Z}} \mid C_{\mathcal{Z}}, \Theta)\, p(C_{\mathcal{Z}} \mid Y_{\mathcal{K}}, C_{\mathcal{K}}, \Theta)\, p(C_{\mathcal{K}}, \Theta \mid Y_{\mathcal{K}})\, dC_{\mathcal{Z}}\, d\Theta\, dC_{\mathcal{K}} \end{aligned} \tag{17}$$

by drawing a $Y_{\mathcal{Z}}$ from $p(Y_{\mathcal{Z}} \,|\, C_{\mathcal{Z}}, \Theta)$ for each sampled value of $C_{\mathcal{Z}}$ drawn from (16). These samples provide full Bayesian inference for all points in $\mathcal{Z}$. If the points in $\mathcal{Z}$ lie within the domain of a cycle, the we obtain the smoothed values of the concentration state and measurements, while if the time points lie outside of the domain (in the future), we obtain forecasting estimates for the concentration state and predictions of measurements based upon values of the explanatory variables in X_t at such points.

5 Simulation

We simulate three experiments. The first generates data from the mechanistic system described in (2) using the parameter values $V = 100$ m^3, $G = 1000$ particles per minute and an average ventilation rate of $Q = 20$ m^3/min. We generated the data from the distribution of Z_t in (11) setting C_t to be the exact solution in (4) and (5) with $C_0 = 10$, $\beta = 0$, $\sigma_v^2 = 0.01$ and $\Delta_t = 1$. We generate only one 20 minute cycle assuming that the particle generator is kept on for the first 15 minutes, which implies $T_0 = 15$ in (4). The second experiment follows the same experimental specifications as the first but simulates 3 cycles three cycles. We assume that the particle generators are kept on for the first 15 minutes within each of the cycles, which implies that $T_0 = 15$ in the mechanistic system (4) for each of the three cycles. We generate the data over 90 observed time points split into $\mathcal{K}_1 = \{1, \ldots, 30\}$, $\mathcal{K}_2 = \{41, \ldots, 70\}$ and $\mathcal{K}_3 = \{81, \ldots, 110\}$.

The third experiment changes the mechanistic alters the mechanistic system from the previous two. Here, we generate data for three cycles from the distribution of Z_t in (11) using the mechanistic system in (6) using $Q_L = Q_R = 5$ m^3/min, $\epsilon_L = \epsilon_{L.F} = 0.5$ and $\epsilon_{R.F} = 0.9$, while retaining the same parameter values for V, G, Q, C_0, β and σ_v^2 as in the first and second experiments. The sets of indices at which data are observed is same as that of the second experiment. We analyze these data using (11); see Section 5.3.

5.1 Priors for mechanistic parameters

Recall that our model parameters are classified into $\Theta_1 = \{\beta, \sigma_v^2\}$ representing parametric linear regression coefficients and a measurement error variance, and $\Theta_2 = \{\phi, m_\omega, \sigma_\omega^2\}$,

where ϕ denotes the parameters in the mechanistic model under consideration. For the most general model in (6), we have $\phi = \{G, Q, Q_R, Q_L, \epsilon_L, \epsilon_{L.F}, \epsilon_{R.F}\}$ while (2) has $\phi = \{G, Q\}$. We use the family of priors specified in (15) with $a_G = 200$, $b_G = 1800$, $a_Q = 3$, $b_Q = 50$, $a_{Q_L} = 2$, $b_{Q_L} = 10$, $a_{Q_R} = 2$, $b_{Q_R} = 10$, $a_{\epsilon_L} = a_{\epsilon_{L.F}} = 0.3$, $b_{\epsilon_L} = b_{\epsilon_{L.F}} = 0.7$, $a_{\epsilon_{R.F}} = 0.6$, $b_{\epsilon_{R.F}} = 1$, $a_\upsilon = 10$, $b_\upsilon = 8.42$, $a_\omega = 2$, $b_\omega = 1.68$, $\mu_m = 0$ and $\kappa_m = 100$.

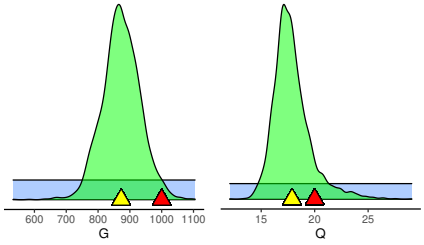
Priors for the set of mechanistic parameters ϕ, which are involved in the process evolution can be defined completely by the user or can be derived from the heuristic methods often followed by the experimenter to get rough estimates of the parameters (see, e.g., the model calibration procedure in Hewett and Ganser, 2017). The methods can include considering the log-transformed concentration only for the decay part of an experiment and regressing them on time. In case of (2), the regression coefficient of time yields estimates of the ventilation rate Q, which, in turn, will provides estimates for G when the log-transformed concentration of the rise in (4) is regressed on time. For more complex models, such as (6), these heuristic methods fail to estimate all the parameters involved with ventilation. Other engineering interventions are necessary to overcome these problems, where they exploit the nested nature of the models. Hewett and Ganser (2017) remarks that calibration procedures are akin to back-of-the-envelope calculations for practicing occupational hygienists. However, these calculations can be used to build reasonable priors for parameters of ϕ.

5.2 Computation

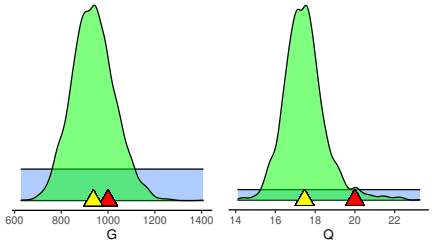
All models discussed here are implemented in `R` 4.2.2 using *rjags* (Plummer, 2022). The posterior inference for each model is based on MCMC chains with 3000 iterations retained after discarding the initial 2000 samples as burn-in. These programs were executed on a single Apple M1 chip, with 3.20 GHz base clock speed and 8 GB of random-access memory running `macOS Ventura` (Version 13.4.1). We assessed convergence of MCMC chains by visually monitoring autocorrelations and checking the coverage of parameter estimates (posterior mean and 95% credible interval) with the true values for the simulated data. Codes and data required to reproduce the results and findings in this article are openly available at Github (active link for downloading).

5.3 Simulation Results

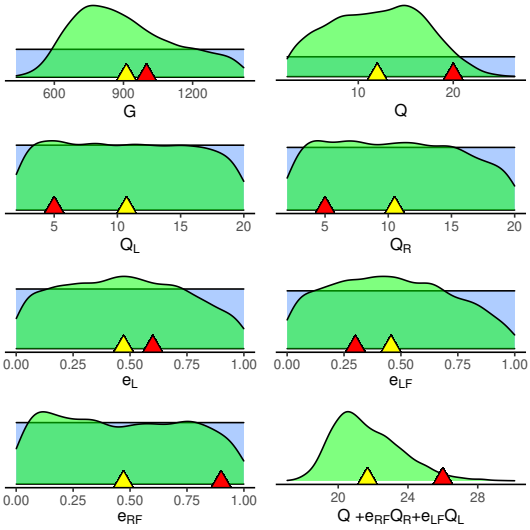
In each of the above simulated experiments, we report data analysis using the hierarchical model (11) with mechanistic systems (2) and 6. For (2), we see reasonable posterior learning for G and Q in Figures 3(a) and 3(b), whereas for (6) the parameters appear to be poorly identifiable. Consequently, we see impaired posterior learning when we assign uninformative priors. Here, the mechanistic parameters $\phi = \{G, Q, Q_R, Q_L, \epsilon_L, \epsilon_{L.F}, \epsilon_{R.F}\}$ appear as functions $(1 - \epsilon_L\epsilon_{L.F})G$ and $(Q + \epsilon_{L.F}Q_L + \epsilon_{R.F}Q_R)$. Therefore, while we see



(a) Mechanistic model in (2) with single cycle experiment



(b) Mechanistic model in (2) with three cycles experiment



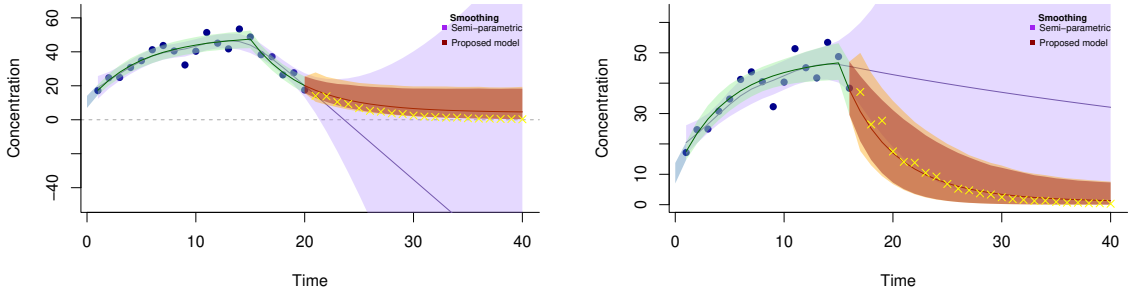
(c) Mechanistic model in (6) with three cycles experiment

Figure 3: Posterior learning for ϕ from mechanistic models in (2) and (6). The red "tick' corresponds to the true value of the parameters in the synthetic experiment, while the yellow "tick" corresponds to their posterior means based on 3,000 samples. The green density shapes the posterior in contrast to the reasonably flat priors shown in blue.

relatively poor learning of individual parameters, learning of the aforementioned functions is reasonable. Hence, learning of the latent process is not compromised. Figure 3(c) depicts reasonable learning for $(Q + \epsilon_{L.F}Q_L + \epsilon_{R.F}Q_R)$ as opposed to weaker learning of the individual parameters Q, Q_R, Q_L, ϵ_L, $\epsilon_{L.F}$ and $\epsilon_{R.F}$. Hence, strongly informative priors are necessary if estimates of these individual parameters are desired.

Investigators studying exposure assessments are interested in estimation of these functional forms instead of individual parameters. For example when modeling dynamics of infectious respiratory aerosols, the quantities Q/V in (2) or $(Q + \epsilon_{L.F}Q_L + \epsilon_{R.F}Q_R)/V$ for (6) correspond to aerosol removal rates that are important in analyzing air changes per hour (ACH), which, in turn, can inform about probability of infection spread.

We also assess the state-space model's effectiveness in capturing the latent process at the observed time points using the posterior predictive distribution (16). Furthermore, for unobserved time points we smooth the latent and observed concentrations using (16) and (17), respectively. Subsequently, we compare the performance of each model with different Bayesian semiparametric regressions that do not incorporate information from the underlying mechanistic system. In particular, we considered methods such as B-splines, natural cubic splines and random walk models of order 2 estimated using Integrated Nested Laplace Approximation (INLA: Rue *and others*, 2009; Lindgren *and others*, 2011). We specifically consider a continuous random walk model on second



(a) Smoothing on untransformed data

(b) Smoothing on log transformed data

Figure 4: Prediction and forecasting performances for our hierarchical model (11) and CRW2-based smoothing on the simulated data for the first few time points (marked in blue) with $T_0 = 15$. In Figures 4(a) and 4(b) we used $t_N = 20$ and $t_N = 16$ respectively.

order increments (denoted CRW2, as described in Section 3.5 of Rue and Held, 2005).

Figure 4 shows a simple out-of-sample analysis comparing our physics-informed model with semiparametric smoothing. The latter, besides delivering wider uncertainty bands, tends to poorly estimate the trajectory compared to the former. The purple line and the band shows the trajectory of the concentrations fitted using semiparametric smoothing along with the uncertainty around it, whereas the red line and the band show the trajectory and associated uncertainty for the out-of-sample points from our proposed physics-informed state-space model. The yellow crosses indicate the out-of-sample data beyond $t_N = 20$ in Figure 4(a) and $t_N = 16$ in Figure 4(b). In Figure 4(a), semiparametric smoothing even forecasts negative concentrations if the data is not appropriately transformed. Figure 4(b) shows that, even under a suitable transformation, forecasts from semiparametric smoothing are sensitive to the time when the data becomes unavailable.

Finally, Table 2 presents overall model comparisons using Watanabe-Akaike Information Criterion (WAIC: Watanabe, 2010, 2013) as implemented in the *LaplacesDemon* (Statisticat and LLC., 2021) package for the `R` statistical computing environment (R Core Team, 2022). We report these scores for the different experimental scenarios and compare (11) with semiparametric smoothing using B-splines, cubic splines, independent second order increments (denoted RW2, as described in Section 3.4 of Rue and Held, 2005) and CRW2. Table 2 shows that while (11) significantly outperforms semiparametric smoothing models in out of sample forecasting, the overall model fit as summarized by WAIC between these methods are much more competitive, and in some cases significantly better, than (11). For the single cycle data, WAIC scores for (11) with the mechanistic model (2) are considerably lower than all other methods, while with (6) all of the models are competitive in a single cycle. On the other hand, the two random walk models produce significantly lower WAIC scores than the others.

Model	Number of cycles	Smoothing by hierarchical model in (11)	Smoothing by Integrated Nested Laplace Approximation					
			B-splines		Cubic splines		Random walk models of order 2	
		WAIC	Knots	WAIC	df	WAIC	Type	WAIC
Model 101	1	-39.1	5	-19.0	3	-16.4	RW2	-19.7
			8	-17.9	10	-18.2	CRW2	-19.4
			20	-12.6	20	-26.2		
	3	-38.8	8	-55.4	7	-12.3	RW2	-121.9
			12	-81.6	15	-85.1	CRW2	-119.6
			24	-10.8.5	20	-96.2		
Model 111	1	-17.2	5	-18.1	3	-16.8	RW2	-19.9
			8	-17.8	10	-17.3	CRW2	-19.5
			20	-12.3	20	-26.8		
	3	12.2	8	8.5	7	52.8	RW2	-128.9
			12	-65.5	15	-82.5	CRW2	-124.9
			24	57.5	20	55.8		

Table 2: Comparison of predictive information criteria between our physics-informed Bayesian state-space models and various Bayesian smoothing techniques using INLA on the simulated data. For B-splines, the knots denotes number of equi-spaced knots for the spline basis. For smoothing using random walks, RW2 model assumes independent second order increments and CRW2 denotes continuous time random walks on second order increments.

That RW2 and CRW2 are excelling in terms of WAIC is likely attributable to their interpolation capabilities surrounding the availability of significantly more data in the 3-cycle experiments. In fact, we see a roughly 21% increase in the residual sum of squares for (11) over RW2. However, we caution against overstating the excellence of these random walk models that have no mechanistic information. As seen in Figure 4, in the absence of mechanistic information forecasting suffers significantly with these random walk models. Furthermore, the aforementioned reduction in the residual sum of squares should warn investigators against over-fitting. Finally, even if these models estimate concentration levels efficiently, they do not inform about the mechanistic process parameters that govern the underlying physics.

6 Analysis of Rail Car Experiment

Das *and others* (2023) have collected substantial concentration data based upon designed experiments with different engineering controls, where each experiment consists of exactly three cycles as presented in Section 3. Here, we do not consider the heterogeneity in ventilation patterns created due to the directional flow of aerosols. Instead, we focus on modeling the data for one experimental run measured at one location inside the rail car.

Due to unavailability of expert prior information on the mechanistic parameters, we only consider the calibration procedures in Hewett and Ganser (2017) to construct priors for the relevant parameters.

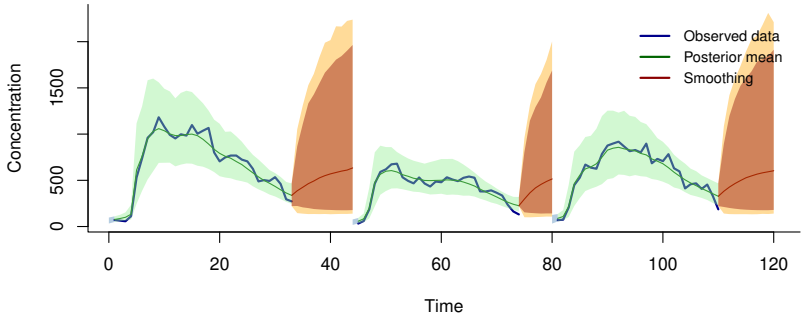
6.1 Noise calibration by mechanistic variance evolution

Experimental data from aerosol concentrations contain considerable amounts of noise, which often exceeds the capabilities of a statistical model equipped with uniform error variance over time to quantify uncertainty. As we expect the aerosol concentration measurements to appear in varying scales across the duration of an experimental cycle, we consider the influence of mechanistic factors on the evolution of the error variance over time. A simple yet effective approach to address this is to introduce a dynamic $v_t = v_t(\phi)$ scale factor in the variance of ω_t in transition equation in (10). This scale factor depends on ϕ since the mechanistic parameters dictate how the data are generated and, hence, how its variability evolves. Therefore, we can modify the transition equation in (11) as

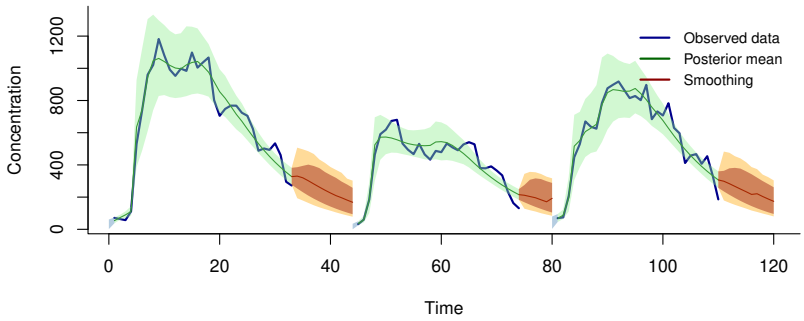
$$\begin{aligned} C_t &= \left(1 - \frac{\Delta}{V}Q\right) C_{t-1} + \frac{\Delta}{V}G_t + v_t\omega_t \\ v_t &= H_t\, v_{t-1} \\ H_t &= (1+\alpha)1_{\mathcal{G}}(t) + \beta(1 - 1_{\mathcal{G}}(t)) \\ \{\alpha, \beta\} &\sim p(\alpha)\, p(\beta)\ , \end{aligned} \tag{18}$$

where $1_{\mathcal{G}}(t) = 1$ when $G_t = G$ (i.e., the particle generator is in place) and $1_{\mathcal{G}}(t) = 0$ when $G_t = 0$ (no generation). With $\alpha > 0$, we model the error variances in the transition equation to change in a multiplicative fashion - increasing as long as the particle generator is on, and decreasing with $0 < \beta < 1$ after the generator is turned off. These modifications are applicable to (8) and (9). Since the process in (9) is derived from a finite difference approximation of the original system, we may not be able to easily characterize the error distribution from a transformation of $\{C_t\}_{t\geq 1}$ in the transition equation while also maintaining an appropriate first-order Markov dependence. The above model is implemented in the computing environment described in Section 5.2. Posterior inference reported here is based on 5,000 MCMC samples after a burn-in of 5000 iterations.

In the current context, we find that our modified (18) adequately provides robust analysis and a similar modification in the observation equation is unnecessary. This choice is corroborated by restrictions on the support of the error variance imposed by $\{C_t\}_{t\geq 1}$. As concentrations are positive quantities, we modeled the transition errors using a log-normal distribution. As a result, when the noisy experimental data is fitted with a model with time-independent transition errors, the parameter estimates in the log-normal distribution yield inaccurate and unreasonably wide uncertainty bands for smoothing and forecasting. Considering time-dependent errors in the transition equation



(a) Time-independent process evolution errors



(b) Time-dependent process evolution errors

Figure 5: Smoothing and forecasting on noisy experimental data by models with and without evolution of error variances over time.

resolves this problem by suitably calibrating the errors informed by the aerosol generation status provided by the mechanistic system. This enriches Bayesian melding of mechanistic information and the statistical model. Figure 5 presents these comparisons.

7 Discussion

Addressing the growing interest among health scientists in assimilating information from physics-based mechanistic systems with experimental data, we undertake such an exercise in an environmental hygiene setting to infer about underlying processes driving transmission of aerosols in closed chambers. Using a Bayesian hierarchical modeling framework, we model the observations from a designed experiment using engineering interventions as a manifestation of latent aerosol concentrations governed by a mechanistic system. Recognizing choices in the statistical approaches to achieve such melding of information, we demonstrate inferential benefits of the generic framework (1) and, more specifically, of state-space models derived using finite-difference approximations of the mechanistic

systems (adapting frameworks outlined, e.g., in Wikle and Hooten, 2010; Abdalla *and others*, 2020, to address specific challenges in the current problem). More specifically, we show that while certain mechanistic parameters may not be well identified by the field data, the latent concentration process is effectively estimated. Assimilating mechanistic systems in our data analysis framework yields especially pronounced benefits in forecasting performance over flexible semiparametric smoothing techniques that do not assimilate such systems, while all these methods may indicate adequate goodness of fit.

Extensions of our models are possible in different directions. For example, in controlled experiments it is typical of photo detector particle counters to offer particle counts in several size ranges that are expected to be correlated. This stokes the possibility of jointly modeling the particle sizes in the process. For p different size-ranges, let Y_t and C_t be p-variate observed and latent concentrations at time t. A multivariate framework for the one-box model we considered is

$$\begin{aligned} g(Y_t) \mid C_t, \beta, \sigma_v^2 &\sim \mathcal{N}_p(g(C_t) + X_t^\top \beta, \Sigma) \\ C_t \mid \phi, m_\omega, \sigma_\omega^2 &\sim \text{ShiftedLMN}(A_t(\phi, \Delta_t)^\top C_{t-1} + B_t(\phi, \Delta_t); m_\omega, \sigma_\omega^2 I_p) \\ \{\phi, \beta, \sigma_v^2, m_\omega, \sigma_\omega^2\} &\sim p(\phi)\, p(\beta \mid \sigma_v^2)\, p(\sigma_v^2)\, p(m_\omega)\, p(\sigma_\omega^2)\ , \end{aligned} \tag{19}$$

where $A_t(\phi, \Delta_t) = 1_p - (Q + \epsilon_{L.F} Q_L + \epsilon_{R.F} Q_R)\Delta_t / V$ and $B_t(\phi, \Delta_t) = (1 - \epsilon_L \epsilon_{L.F}) G 1_{\mathcal{G}}(t) \Delta_t / V$ are calculated using element wise operations applied to the p-variate parameters in ϕ. Here, $X + \theta$ is distributed as $\text{ShiftedLMN}(\theta; \mu, V)$ if $\log X$ is distributed as multivariate normal with mean μ and covariance matrix V for some $\theta \in \mathbb{R}^p$. Further investigations into the dependence structure among size-specific particle concentrations is open to future investigations as are questions on the structure of Σ in (19) and its effects of inference.

While the current analysis advocates delving in the mechanistic equations as a part of the model building exercise, we recognize that such luxuries may be precluded by more complex models in other applications. In this regard, stochastic emulators such as Gaussian processes are widely employed to conduct such inference. We have not undertaken a comprehensive comparison with such methods in this paper and recognize them as viable options in our current setting. This comprises an area of future research.

Acknowledgments

Conflict of Interest: None declared.

Funding

Banerjee, Ramachandran and Pan were supported, in part, from the National Institute of Environmental Health Sciences (NIEHS) R01ES030210. Banerjee also acknowledges

support from NIEHS R01ES027027, the National Institute of General Medical Science (NIGMS) R01GM148761, and the Division of Mathematical Sciences (DMS) of the National Science Foundation 2113778.

References

Abdalla, Nada, Banerjee, Sudipto, Ramachandran, Gurumurthy and Arnold, Susan. (2020). Bayesian state space modeling of physical processes in industrial hygiene. *Technometrics* **62**(2), 147–160.

Bazant, Martin Z. and Bush, John W. M. (2021). A guideline to limit indoor airborne transmission of covid-19. *Proceedings of the National Academy of Sciences* **118**(17), e2018995118.

Chia, Po Ying, Coleman, Kristen Kelli, Tan, Yian Kim, Ong, Sean Wei Xiang, Gum, Marcus, Lau, Sok Kiang, Lim, Xiao Fang, Lim, Ai Sim, Sutjipto, Stephanie, Lee, Pei Hua, Son, Than The, Young, Barnaby Edward, Milton, Donald K., Gray, Gregory C., Schuster, Stephan, Barkham, Timothy, De, Partha Pratim, Vasoo, Shawn, Chan, Monica, Ang, Brenda Sze Peng, Tan, Boon Huan, Leo, Yee-Sin, Ng, Oon-Tek, Wong, Michelle Su Yen, Marimuthu, Kalisvar, Lye, David Chien, Lim, Poh Lian, Lee, Cheng Chuan, Ling, Li Min, Lee, Lawrence, Lee, Tau Hong, Wong, Chen Seong, Sadarangani, Sapna, Lin, Ray Junhao, Ng, Deborah Hee Ling, Sadasiv, Mucheli, Yeo, Tsin Wen, Choy, Chiaw Yee, Tan, Glorijoy Shi En, Dimatatac, Frederico, Santos, Isais Florante, Go, Chi Jong, Chan, Yu Kit, Tay, Jun Yang, Tan, Jackie Yu-Ling, Pandit, Nihar, Ho, Benjamin Choon Heng, Mendis, Shehara, Chen, Yuan Yi Constance, Abdad, Mohammad Yazid, Moses, Daniela *and others*. (2020, May). Detection of air and surface contamination by sars-cov-2 in hospital rooms of infected patients. *Nature Communications* **11**(1), 2800.

Das, Darpan, Babik, Kelsey R., Moynihan, Emma and Ramachandran, Gurumurthy. (2023). Experimental studies of particle removal and probability of covid-19 infection in passenger railcars. *Journal of Occupational and Environmental Hygiene* **20**(1), 1–13. PMID: 36256520.

Fuentes, Montserrat and Raftery, Adrian E. (2005, March). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics* **61**, 36–45.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki and Rubin, Donald B. (2013). *Bayesian Data Analysis, 3rd Edition*, Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC.

Harris, Jeffrey E. (2020, April). The subways seeded the massive coronavirus epidemic in new york city. *Working Paper* 27021, National Bureau of Economic Research.

Hewett, Paul and Ganser, Gary H. (2017). Models for nearly every occasion: Part i - one box models. *Journal of Occupational and Environmental Hygiene* **14**(1), 49–57. PMID: 27869546.

Lindgren, Finn, Rue, Håvard and Lindström, Johan. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society B* **73**(4), 423–498.

Monteiro, J V. D., Banerjee, Sudipto and Ramachandran, Gurumurthy. (2014). Bayesian modeling for physical processes in industrial hygiene using misaligned workplace data. *Technometrics* **56**(2), 238–247.

New York City Metropolitan Transportation Authority. (2020). Subway and bus ridership for 2020. [Accessed Jun 11, 2021]. https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2020.

Plummer, Martyn. (2022). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-13.

Poole, David and Raftery, Adrian E. (2000, Dec). Inference for deterministic simulation models: The bayesian melding approach. *Journal of the American Statistical Association* **95**, 1244–1255.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, Adrian E. and Bao, Le. (2010*a*, Dec). Estimating and projecting trends in hiv/aids generalized epidemics using incremental mixture importance sampling. *Biometrics* **66**, 1162–1173.

Raftery, Adrian E. and Bao, Le. (2010*b*, Dec). Estimating and projecting trends in hiv/aids generalized epidemics using incremental mixture importance sampling. *Biometrics* **66**, 1162–1173.

Raftery, Adrian E., Givens, Geof H. and Zeh, Judith E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association* **90**, 402–416.

Reinke, P.H. and Keil, C.B. (2009). *Mathematical Models for Estimating Occupational Exposure to Chemicals*. American Industrial Hygiene Association.

Rue, Havard and Held, Leonhard. (2005). *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*, Chapter 3. Chapman & Hall/CRC, pp. 118–140.

Rue, Håvard, Martino, Sara and Chopin, Nicholas. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B* **71**, 319–392.

Statisticat and LLC. (2021). *LaplacesDemon: Complete Environment for Bayesian Inference*. R package version 16.1.6.

Watanabe, Sumio. (2010, dec). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594.

Watanabe, Sumio. (2013, mar). A widely applicable bayesian information criterion. *J. Mach. Learn. Res.* **14**(1), 867–897.

Wikle, Christopher and Hooten, Mevin. (2010). A general science-based framework for dynamical spatio-temporal models. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* **19**(3), 417–451.

Wikle, Christopher K., Zammit-Mangion, Andrew and Cressie, Noel. (2019). *Spatio-Temporal Statistics with R*. Boca Raton, FL: Chapman & Hall/CRC.

Zhang, Yufen, Banerjee, Sudipto, Yang, Rui, Lungu, Claudiu and Ramachandran, Gurumurthy. (2009). Bayesian modeling of exposure and air flow using two-zone models. *The Annals of Occupational Hygiene* **53**(4), 409–424.